NACSIS Test Collection Workshop (NTCIR-1)

Noriko Kando, Kazuko Kuriyama, Toshihiko Nozue
Research and Development Department
National Center for Science Information Systems (NACSIS) , Tokyo 112-8640, Japan
URL: http://www.rd.nacsis.ac.jp/~{kando,ntcadm}/

1. Introduction

This paper reports the outline of the first NACSIS Test Collection for Information Retrieval (NTCIR) Workshop[1] will be held on August 30-September 1, 1999[1]. It is the first competition-type workshop designed to enhance research in Japanese text retrieval.

The NTCIR Workshop has the following goals;

(a) To encourage research in information retrieval, cross-lingual information retrieval and related areas by providing a large-scale Japanese test collection and a common evaluation setting that allows cross-system comparisons

(b) To provide a forum for research groups interested in comparing results and exchanging ideas or opinions in an informal atmosphere

(c) To improve the quality of the Test Collections based on the feedback from participants

The test collection used in the Workshop consists of more than 330,000 documents and more than half are English-Japanese paired. Although there is a Japanese test collection called BMIR-J2 consisting of 5,080 newspaper articles[2], enhancement of the Japanese test collection in the both aspects of the variety of text types and the scale is needed. We put emphasis on cross-lingual retrieval since it is critical in the internet environment and Japanese scientific information retrieval [3].

Thirty-one groups including participants from six countries had enrolled. Among them, twenty-eight groups enrolled in IR tasks (twenty-three in Ad Hoc task and sixteen in Cross-lingual task), and ten in Term Recognition task. Ten are from companies and twenty-one are from universities or national research institutes. Nineteen groups are Japanese, eight are non-Japanese and four are mixed. No company participants from outside of Japan.

In the next section, we describe the tasks performed in the Workshop. Section 3 shows the test collection (NTCIR-1) used in the Workshop. The final section lists issues to be discussed.

2. The Tasks

A participant conducted one or more of the tasks below:

(a) **The Ad Hoc Information Retrieval task :** to investigate the retrieval performance of systems that search a static set of documents using new search topics

(b) **The Cross-Lingual Information Retrieval task :** an ad hoc task in which documents are in English and topics are in Japanese.

(c) **The Automatic Term Recognition and Roll Analysis task :** (1) to extract terms from titles and abstracts, and (2) to identify the terms representing the "object", "method" and "main operation" of the main topic.

2.1 The Procedures

In November, 1998, the document data, thirty ad hoc topics, twenty-one cross-lingual topics and their relevance assessments were provided for each IR tasks participant to train their systems. The 53 new test topics were distributed on Feb. 8 and the search results for them were submitted by March 4 as official test runs. The test topics are common for both IR tasks.

A participant can submit the results of more than one run. Both automatic and manual query constructions are allowed. In the case of automatic construction, the participants must submit at least one set of results of the searches using only "DESCRIPTION" fields of the topics as the mandatory runs. For optional automatic runs and manual runs, any fields of the topics can be used. Also each participant had to fill and submit a system description form describing the detailed feature of the system.

Human analysts assess the relevance of retrieved documents to each topic. Based on the relevant assessments, inter-polated recall and precision at 11 points, average precision (non-interpolated) over all relevant documents, and precisions at 5, 10, 15, 20, 30, 100 documents will be calculated using TRECs evaluation program, which is available from the ftp site of Cornell University.

3. The Test Collection

The test collection used in the Workshop is consists of; documents, topics, and relevance assessments for each search topic.

3.1 Documents

The documents are author abstracts of the papers with wide range of length and are presented at academic meetings hosted by 65 Japanese academic societies. Since one of the purposes of the original database is to provide an alert information about papers presented in Japanese academic conferences as soon as possible, documents are put in the database without any revision nor modification by professional abstractors or editors. Some of them are refereed, and others are pre- or non-refereed.

The Collection contains three document collections, *i.e.* JE Collection, J Collection, and E Collection. JE Collection contains 339,483 documents, more than half are English-Japanese paired. J and E Collections are constructed through extracting Japanese or

[1] This project is supported by “Research for the Future” Program JSPS-RFTF96P00602 of the Japan Society for the Promotion of Science

English parts of the documents, respectively, from the JE Collection

In the Workshop JE Collection is used in the Ad Hoc task since Japanese operational IR environment, especially, retrieval of scientific documents and Web documents, retrieving both Japanese and English documents at a time is quite natural. E Collection is used in the Cross-lingual task. J Collection is used in the monolingual retrieval, which will be the baseline for comparing the search effectiveness with the results in the Cross-lingual task.

Documents are SGML tagged plain text. A record may contain document ID, title, a list of author(s), name and date of the conference. abstract, keyword(s), and name of the hosted society.

3.2 Topics

A topic is a formatted description of a user's need. We defined the topics as statements of "user need" rather than "queries" which are the strings actually submitted into the system since we would like to allow both manual and automatic query construction.

Its format is similar to the one once used in the TREC-1 and 2 and contains SGML-like tags. A topic consists of a title of the topic, a description, a detailed narrative, a list of concepts and field(s). The title is a very short description of the topic and can be used as a very short query which resembles the one often submitted by an end-user of internet search engines. Each narrative may contain detailed explanation of the topic, term definition, background knowledge, purpose of the search, criteria of relevance judgement, and so on.

3.2.1 Topic Preparation

Some topics are collected from users with permissions to use as part of a test collection, and others are created by the analysts. Each topic was examined its clearness and difficulty. We selected topics with more than five relevant documents in test searches done in NACSIS. Sentences were modified when they were too restricted or ambiguous. We had tried to balance the topic length, number of relevant documents, and "difficulty" however we found that estimating "difficulty" is difficult.

3.3 Relevance assessments (Right Answers)

The relevance assessment is done in three grades, i.e., relevant, partially relevant, non-relevant. Two analysts assess the relevance of a topic separately, then the primary analyst of the topic who created the topic decides the final judgment.

3.3.1 Relevance assessments for Training Topics

Regarding training topics, the top ranked documents in the search results of various search strategies using three IR systems in NACSIS and recall-oriented manual searches are pooled and formed a set of candidates of relevant documents.

3.3.2 Analysis of Pooling in the Pre-test

On Dec. 2, as a pre-test, top1000 documents for each training topics were asked to submitted (1) to obtain the feedback for the training topics and the relevance assessments for them, and (2) to confirm the procedure of the test. By adding additional relevant documents found in the pretest to the initial ones, the relevance assessments for training topics were revised.

Using the results of the pretest, we evaluated (1) coverage of the initial pooling done in NACSIS, (2) effectiveness of pooling, and (3) the reliability of a test collection through investigating the effect of variations of pooling methods and coverage, and variation in relevance assessments have on the evaluation of search effectiveness. The results are as follows;

(1) The initial pooling worked well and covered 97% of the total relevant documents.

(2) Interactive searches was effective for some particular topics and found 17.5% of unique relevant documents.

(3) In the aspects of exhaustivity, the pooling top 100 documents from each run worked well for the topics with less than 50 relevant documents. As for the topics with more than 100 relevant documents, though top 100 pooling covered only 51.9% of the total relevant documents, the coverage reached to 90% if it was combined with interactive searches.

(4) We found very high similarity among the rankings of systems produced using different sets of relevance assessments regardless of the different coverage and pooling methods and regardless of inconsistency among relevance assessments.

(5) The top100 pooling method has an effect to reduce the size of the document pooling to 27% of the possible size of it.

Regarding test topics, based on the analyses above we decided to use top100 pooling.

3.4. Linguistic Analysis

A part of the J collection contains detailed part-of-speech tags [4]. Because of absence of explicit boundary between words in Japanese sentences, we set the three levels of lexical boundaries (i.e., word boundary, strong and week morpheme boundary), and assigned detailed POS tags based on the boundaries and types of origin, so that the collection can be used to examine the suitable term segmentation of Japanese texts for retrieval purpose.

4. Summary and Future Directions

For the official runs, 45 ad hoc runs and 66 cross-lingual runs were submitted. Relevance assessment is now going on and is expected to be finished by the end of May. The followings are the issues we have to consider to further enhancement;

- Schedule: Is the training period sufficient?
- Is the support regarding Japanese text processing sufficient for non-Japanese participants?
- Is a smaller collection also needed? Some participants could not complete because of the size of the collection.
- Evaluation method of the difficulty of search topics
- Enhance the variation of text types
- International collaboration for cross-lingual collections

ACKNOWLEDGMENTS

We thanks for all the participants for their contributions and analysts who worked very hard with surprisingly excellent concentration. Our special thanks due to Donna Harman, Ellen Voorhees, Ross Wilkinson, Sung H Myaeng, and MunKew Leong for their substantial advise and continuous support.

REFERENCES

[1] NTCIR Project. http://www.rd.nacsis.ac.jp/~ntcadm/

[2] Kando, N. Cross-linguistic scholarly information transfer and database services in Japan. Annual Meeting of the ASIS. (Nov. 1997) Washington DC.

[3] Kando, N. et al. NTCIR: NACSIS Test Collection Project [poster]. the 20th Annual BCS-IRSG Colloquium, (March 1998), Autrans, France.

[4] Kageura, K. et al. NACSIS Corpus Project for IR and Terminological Research. NLPPR '97, (Dec. 1997), 493-496.